Current state of the publicly available cyanobacteria sequencing database Analysis

Yilin Qiu

1. Downloading raw sequencing reads from NCBI

- PacBio datasets: Downloading 85 raw sequencing datasets from NCBI

  Directory of the raw reads:

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.fastq`

- Illumina paried-end datasets: Downloading 458 PacBio raw sequencing datasets from NCBI

  Directory of the raw reads:

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*.fastq.gz`

2. Trimming the raw sequencing reads:

- PacBio datasets:

  Flye assembler take raw PacBio sequencing reads. Trimming step is skipped for PacBio datasets.

- Illumina paried-end datasets: Trimming the raw sequencing reads by metaWRAP - fastqc - v.0.11.8:

```
metawrap read_qc -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_2.fastq -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/
```

```
metawrap read_qc -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_2.fastq -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/
```

```
metawrap read_qc -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_458_sep7_paried_fastq/*_2.fastq -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/
```

Directory of the trimmed sequencing reads:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/*.fastq`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.fastq`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.fastq`

3. Assembly raw sequencing reads:

- PacBio datasets: Assembly using flye assembler - v. 2.8.3 with default parameters:

```
  /home/yqiu/miniconda2/envs/flye2.8/bin/flye --pacbio-raw
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.fastq --out-dir *.flye --
meta
```

  Directory of the assemblies:

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.flye/assembly.fasta`

- Illumina paried-end datasets: Assembly by MegaHIT - v. 1.1.3 with default parameters:

```
metawrap assembly -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/*_2.fastq -m 10 -t
10 --megahit -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*
```

```
metawrap assembly -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*_2.fastq -m 10 -t
10 --megahit -o /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*
```

```
metawrap assembly -1
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*_1.fastq -2
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*_2.fastq -m 10 -t 10 --
megahit -o /mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.fasta
```

  Directory of the assemblies:

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*/*.fasta`

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*.fasta`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.fasta`

4. MetaWRAP binning - Maxbin2 -v.2.2.6

- PacBio datasets:
  - Maxbin2 bining with default parameters, Min contig length: 1000:

    ```
    metawrap binning -o metawrap-binning -t 20 -a assembly.fasta --maxbin2 --single-end /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.fastq -o Tripe_binning/maxbin2_bins
    ```

    Directory of the bins:

    `/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.flye/Tripe_binning/maxbin2_bins`

  - Quality assessment for the bins by Checkm_DB with default parameters:

    ```
    metawrap bin_refinement -A /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.flye/Tripe_binning/maxbin2_bins -t 12 -c 0 -x 100 -o bin_refinemnt2
    ```

    Directory of the assessment for the bins:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.flye/bin_refinemnt2/binsA.stats`

- Illumina paried-end datasets:
  - Maxbin2 bining with default parameters, Min contig length: 1000:

    ```
    metawrap binning -a /mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*/*.fasta -o /mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/* -t 12 --maxbin2 /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/*_1.fastq /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim/*/*_2.fastq
    ```

    ```
    metawrap binning -a /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.fasta -o /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/maxbin2_bins -t 12 --maxbin2 /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*_1.fastq /mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*_2.fastq
    ```

```
metawrap binning -a
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.fasta -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/maxbin2_bins -t 12 --
maxbin2 /mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*_1.fastq
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*_2.fastq
```

Directory of the bins:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/maxbin2_bins/`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/maxbin2_bins/`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/maxbin2_bins`

- Quality assessment for the bins by Checkm_DB with default parameters:

```
metawrap bin_refinement -A
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/ma
xbin2_bins -t 12 -c 0 -x 100 -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/Re
finemnt_maxbin2
```

```
metawrap bin_refinement -A
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/max
bin2_bins -t 12 -c 0 -x 100 -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/Bin
_refinement
```

```
metawrap bin_refinement -A
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/maxbin2_bins -t 12 -c
0 -x 100 -o
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/Refinemnt_maxbin2
```

Directory of the assessment for the bins:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/Refinemnt_maxbin2/binsA.stats`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/Bin_refinement/binsA.stats`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/Refinemnt_maxbin2/binsA.stats`

5. Taxonomy prediction - GTDB-Tk - v.0.3.2

- PacBio datasets: GTDB-Tk taxonomy prediction with default parameters:

```
export GTDBTK_DATA_PATH=/usr/local/gtdbtk_data
gtdbtk classify_wf --genome_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_pacbio_84_sep_7/*.flye/Tripe_binning/maxbin2_
bins --out_dir gtdbtk -x fa
```

  Directory of the results:

  `/home/yqiu/ncbi_pacbio/*.flye/gtdbtk/`

- Illumina paried-end datasets: GTDB-Tk taxonomy prediction with default parameters:

```
export GTDBTK_DATA_PATH=/usr/local/gtdbtk_data
gtdbtk classify_wf --genome_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/maxbin
2_bins --out_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/*.gtdb
tk -x fa
```

```
export GTDBTK_DATA_PATH=/usr/local/gtdbtk_data
gtdbtk classify_wf --genome_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/Triple_bin/maxbin2
_bins/ --out_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.gtdbtk
```

```
export GTDBTK_DATA_PATH=/usr/local/gtdbtk_data
gtdbtk classify_wf --genome_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/maxbin2_bins --out_dir
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.gtdbtk
```

  Directory of the results:

  `/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_binning_maxbin2_meatabat_1/*/*.gtdbtk`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.gtdbtk`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.gtdbtk`

6. Plasmids identifiction - PlasFlow -v.1.1

- PacBio datasets and Illumina paried-end datasets:

  Prediction of plasmid sequences in cyanobacteria genomes by PlasFlow with default parameters:

```
PlasFlow.py --input
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow/*.fa --output
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow/*.plasflow.tsv
```

```
PlasFlow.py --input
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow_2/*.fa --output
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow_2/*.plasflow.tsv
```

  Directory of the results:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow/*.plasflow.tsv`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumina_cyano_plasflow/*.plasflow.tsv`

7. Viral contigs identifiction - VirSorter2 - v. 1.0.6

- PacBio datasets: Virus identification by VirSorter2 with default parameters:

```
virsorter run -i  /media/nfs/data/sandbox/yilinqiu/ncbi_pacbio_84_sep_7/*.flye/*.fasta
-d /home/yilinqiu/vs2/db/ -w /home/yilinqiu/ncbi_pacbio/virsorter.*.out -j 12
```

  Directory of the results:

`/home/yilinqiu/ncbi_pacbio/virsorter.*.out`

- Illumina paried-end datasets: Virus identification by VirSorter2 with default parameters:

```
virsorter run -i
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*/*.fasta -d -d
/mnt/nfs/sharknado/Sandbox/Yilin/db/db_vs2/db -w
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*/*.virsorter.ou
t -j 12
```

```
virsorter run -i
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.fasta -d
/mnt/nfs/sharknado/Sandbox/Yilin/db/db_vs2/db -w
/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.out -j 12
```

```
virsorter run -i
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/final_assembly.fasta -d
/mnt/nfs/sharknado/Sandbox/Yilin/db/db_vs2/db -w
/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.virsorter.out -j 12
```

Directory of the results:

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumia_paried_assembly_megahit_1/*/*.virsorter.out`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/ncbi_illumia_400_after_trim_2/*/*.out`

`/mnt/nfs/sharknado/Sandbox/Yilin/db/illumina_400_leave_out/*/*.virsorter.out`